DOCUMENT RESUME

ED 272 577 TM 860 493

AUTHOR Holland, Paul W.; Thayer, Dorothy T.
TITLE Differential Item Performance and the Mantel-Haenszel Procedure.
PUB DATE Apr 86
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Comparative Analysis; Hypothesis Testing; *Item Analysis; *Latent Trait Theory; Mathematical Formulas; *Response Style (Tests); Statistical Analysis; *Test Bias; Test Items
IDENTIFIERS Chi Square; *Differential Item Performance; *Mantel Haenszel Procedure

ABSTRACT

The Mantel-Haenszel procedure (MH) is a practical, inexpensive, and powerful way to detect test items that function differently in two groups of examinees. MH is a natural outgrowth of previously suggested chi square methods, and it is also related to methods based on item response theory. The study of items that function differently for two groups of examinees, originally called item bias research, focuses on the fact that different groups of examinees may react differently to the same test question. The preferred, more neutral, term is differential item functioning (DIF). In studying DIF, members of the focal group and the reference group should be comparable. The item data may be arranged into 2 x 2 tables. Chi square procedures test a hypothesis, but do not produce a parametric measure of the amount of DIF exhibited by the studied item. The MH chi square test provides a measure of the size of the departure of the data from the null hypothesis. The data are presented in 2 x 2 tables, and the measure of DIF is in the scale of differences in item difficulty as measured in Educational Testing Services delta scale. MH is significantly less expensive to use than item response analyses. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED272577

Differential Item Performance and the Mantel-Haenszel Procedure

by

Paul W. Holland

and

Dorothy T. Thayer

Research Statistics Group
Educational Testing Service
Princeton, NJ 08541-0001

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

D. H Urban

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the American Educational Research Association Annual Meeting, San Francisco, California, April 1986. (Revised April 22, 1986)

TM860493

1. INTRODUCTION AND NOTATION

Holland (1985) proposed the use of the Mantel-Haenszel procedure as a practical and powerful way to detect test items that function differently in two groups of examinees. In this paper we show how this use of the Mantel-Haenszel (MH) procedure is a natural outgrowth of the previously suggested chi-square procedures of Scheuneman (1979), Marascuilo and Slaughter (1981), Mellenberg (1982), and others and we show how the MH procedure relates to methods based on item response theory, Lord (1980).

The study of items that function differently for two groups of examinees has a long history. Originally called "item bias" research, modern approaches focus on the fact that different groups of examinees may react differently to the same test question. These differences are worth exploring since they may shed light both on the test question and on the experiences and backgrounds of the different groups of examinees. We prefer the more neutral terms, differential item performance or differential item functioning, (i.e., dif), to item bias since in many examples of items that exhibit dif the term "bias" does not accurately describe the situation.

Early work at ETS on dif began with Cardall and Coffman (1964) and Angoff and Ford (1973). The book by Berk (1982) summarizes research to 1980.

The following notational scheme and terminology is used in the rest of this paper. We will always be comparing two groups of examinees, of which the performance of one, the focal group, F, is of primary interest. The performance of the other group, the reference group, R, is taken as a standard against which we will compare the performance of the focal group. For example, the focal group might be all black examinees while the reference group might consist of the white examinees. Typically, all test items in a given testing instrument will

be analysed for evidence of dif, and this will be done one item at a time. We will refer to the item that is being examined for evidence of dif in a given analysis as the studied item.

Basic to all modern approaches to the study of dif is the notion of comparing only comparable members of F and R in attempting to identify items that exhibit dif. Comparability means identity in those measured characteristics in which examinees may differ and that are strongly related to performance on the studied item. Important among the criteria used to define comparability are (a) measures of the ability for which the item is designed, (b) schooling or other measures of relevant experience, and (c) membership in other groups. In practice, the matching criteria will usually include test scores since these are available, accurately measured, and usually measure the same ability as the studied item.

If both examinee ability and item characteristics are confounded by simply measuring the difference in the performance on an item between unmatched reference and focal group members, the result is a measure of impact rather than of differential item performance. For example, comparing the proportion of reference and focal group members who give correct answers to a given item is a measure of the item's impact on the focal group relative to the reference group. In this paper we do not discuss impact, since the confounding of differences in examinee ability with characteristics of items is of little utililty in attempting to identify items that may truely disadvantage some subpopulations of examinees.

Suppose that criteria for matching have been selected, then the data for the studied item for the examinees in R and F may be arranged into a series of 2×2 tables; one such table for each matched set of reference and focal group members. The data for the performance of the j^{th} matched set on the studied

item is displayed below

		Score on Studied Item		
		1	0	Total
Group	R	A_j	B_j	n_{Rj}
	F	C_j	D_j	n_{Fj}
	Total	m_{1j}	m_{0j}	T_j

Table 1: Data for the $j^{\underline{th}}$ matched set of members of R and F.

In Table 1, T_j is the total number of reference and focal group members in the $j^{\underline{th}}$ matched set; n_{Rj} is the number of these who are in R; and of these A_j answered the studied item correctly. The other entries in Table 1 have similar definitions.

In order to state statistical hypotheses precisely, it is necessary to have a sampling model for the data in Table 1. It is customary to act as though the values of the marginal totals, n_{Rj} and n_{Fj}, are fixed and to regard the data for R and F as having arisen as random samples of size n_{Rj} and n_{Fj} from large matched pools of reference and focal group members. It follows that A_j and C_j are independent binomial variates with parameters (n_{Rj}, p_{Rj}) and (n_{Fj}, p_{Fj}), respectively. These population values can be arranged as a 2×2 table that is parallel to Table 1; i.e.,

		Score on Studied Item		
		1	0	Total
Group	R	p_{Rj}	q_{Rj}	1
	F	p_{Fj}	q_{Fj}	1

Table 2: Population parameters for data from the $j^{\underline{th}}$ matched set.

The hypothesis of no dif corresponds to the null hypothesis.

$$H_0 : p_{Rj} = p_{Fj} \quad \text{for all } j.$$

The hypothesis, H_0, is also the hypothesis of conditional independence of group membership and the score on the studied item given the matching variable (Bishop, Fienberg, and Holland, 1975).

Under H_0, the "expected values" for the cell entries of Table 1 are well-known to be obtained by the "product of margins over total" rule and are summarized below

$$E(A_j) = n_{Rj}\, m_{1j}/T_j \qquad E(C_j) = n_{Fj}\, m_{1j}/T_j$$

$$E(B_j) = n_{Rj}\, m_{0j}/T_j \qquad E(D_j) = n_{Fj}\, m_{0j}/T_j. \qquad (1)$$

2. PREVIOUS CHI-SQUARE PROCEDURES

Scheuneman (1979) proposed a procedure to test the hypothesis, H_0, utilizing a specific type of matching criterion. Let S denote a score on a criterion test -- e.g., an operational test score that may or may not include the studied item. The values of S are categorized into a few intervals -- Scheuneman suggests that three to five intervals are satisfactory. The matched groups are defined by the categorized values of S so that members of R and F are considered matched if their scores on S fall into the same score interval. In terms of the notation of section 1, the test statistic proposed by Scheuneman is given by

$$\text{SCHEUN} = \sum_{j=1}^{K} \left[\frac{(A_j - E(A_j))^2}{E(A_j)} + \frac{(C_j - E(C_j))^2}{E(C_j)} \right], \qquad (2)$$

which is algebraically equal to

$$\text{SCHEUN} = \sum_{j=1}^{K} \left[\frac{(A_j - E(A_j))^2}{n_{Rj}\, n_{Fj}\, m_{1j}/T_j^2} \right].$$

It was originally thought that SCHEUN had an approximate chi-square distribution on K-1 degrees of freedom when H_0 is true, Schueneman (1979). This is not correct as discussed in Barker (1981) and Scheuneman (1981). For example, under H_0, the expectation of SCHEUN, conditional on the four marginal values n_{Rj}, n_{Fj}, m_{1j}, and m_{0j} in each 2×2 table, is given by

$$E(\text{SCHEUN}) = \sum_{j=1}^{K} \frac{m_{0j}}{(T_j-1)} . \tag{3}$$

This value is sensitive to the total number of incorrect responses in each 2×2 table and can range from 0 up to K. If SCHEUN had an approximate chi-square distribution on K-1 degrees of freedom then the expected value in (3) would be approximately K-1 for any set of values of m_{0j}. Fortunately, a small correction to (2) does give the resulting statistic an approximate chi-square distribution under H_0. The corrected statistic is

$$\text{CHISQ-P} = \sum_{j=1}^{K} \frac{T_j}{m_{0j}} \left[\frac{(A_j - E(A_j))^2}{E(A_j)} + \frac{(C_j - E(C_j))^2}{E(C_j)} \right] \tag{4}$$

which can be shown to be algebraically identical to

$$\text{CHISQ-P} = \sum_{j=1}^{K} \left[\frac{(A_j - E(A_j))^2}{n_{Rj}\, n_{Fj}\, m_{0j}\, m_{1j}/T_j^3} \right] . \tag{5}$$

This is well-known to be the Pearson chi-square test statistic for testing H_0 and the proper degrees of freedom equals the number of matched groups, K, if the T_j are <u>all</u> large, Bishop, Fienberg, and Holland (1975). It is also called the "full" chi-square by some to distinguish it from SCHEUN.

The K 2×2 tables may be regarded as a single 2×2×K table and the standard theory of log-linear models for three-way tables may be used to test H_0. This

leads to the suggestion of Marascuilo and Slaughter (1981) and Mellenberg (1982) to use the likelihood ratio chi-square statistics to test H_0 instead of (5).

The alternative hypothesis against which H_0 is tested by CHISQ-P (and its likelihood-ratio versions) is simply the negation of H_0, i.e.,

$$\bar{H}_0 : p_{Rj} \neq p_{Fj} \quad \text{for some } j.$$

This is why CHISQ-P is a multi-degree of freedom chi-square test. It is not powerful against specific alternatives to H_0, but it will detect <u>any</u> such departure if the T_j are large enough. This fact leads to a trade-off between bias and statistical power that is not well made, in our opinion, by procedures like Scheuneman's or "methods 1, 2, 4, 5, and 6" of Marascuilo and Slaughter (1981). The trade-off arises by the desire to increase the values of T_j in order to increase the power of the test. This degrades the quality of the matching (i.e., lumps together examinees whose scores are not equal) in order to increase the sample sizes in the matched groups, i.e., T_j. This is necessary in these procedures because of the goal of being able to detect <u>any</u> type of departure from H_0. An alternative approach, and one that we favor, is to reduce the types of alternatives to H_0 against which the test has good power and to concentrate this power into a few degrees of freedom that actually occur in test data. This occurs in Method 3 of Marascuilo and Slaughter (1981). Mellenberg (1982) has moved in this direction by distinguishing "uniform" from "non-uniform bias." The M-H procedure does this by concentrating on Mellenberg's uniform bias and yet it does not degrade the quality of the matching. We will discuss this in the next section.

A separate problem with the chi-square procedures is that they are only <u>tests</u> of H_0 and do not produce a parametric measure of the amount of <u>dif</u> exhibited by the studied item. As is well-known, tests will always reject the

null hypothesis provided that the relevant sample sizes are large enough. It is more informative to have a measure of the size of the departure of the data from H_0. The M-H procedure provides such a measure.

3. THE MANTEL-HAENSZEL PROCEDURE

In their seminal paper, Mantel and Haenszel (1959) introduced a new procedure for the study of matched groups. The data are in the form of K 2×2 tables as in Table 1. They developed a chi-square test of H_0 against the specific alternative hypothesis

$$H_1 : \frac{P_{Rj}}{q_{Rj}} = \alpha \frac{P_{Fj}}{q_{Fj}} \qquad j = 1,\ldots,K \tag{6}$$

for $\alpha \neq 1$. Note that $\alpha=1$ corresponds to H_0, which can also be expressed as:

$$H_0 : \frac{P_{Rj}}{q_{Rj}} = \frac{P_{Fj}}{q_{Fj}} \qquad j = 1,\ldots,K. \tag{7}$$

The parameter α is called the common odds-ratio in the K 2×2 table because under H_1, the value of α is the odds ratio

$$\alpha = \frac{P_{Rj}}{q_{Rj}} \Big/ \frac{P_{Fj}}{q_{Fj}} = \frac{P_{Rj}\, q_{Fj}}{P_{Fj}\, q_{Rj}} \quad \text{for all } j = 1,\ldots,K. \tag{8}$$

The Mantel-Haenszel chi-square test statistic is based on

$$\frac{\left(\sum_j (A_j - \sum_j E(A_j))\right)^2}{\Sigma\, \mathrm{Var}(A_j)} \tag{9}$$

where $E(A_j)$ is defined in (1) and

$$\mathrm{Var}(A_j) = \frac{n_{Rj}\, n_{Fj}\, m_{1j}\, m_{0j}}{T_j^2 (T_j - 1)} . \tag{10}$$

The statistic in (9) is usually given a continuity connection to improve the

accuracy of the chi-square percentage points as approximations to the observed significance levels. This has the form

$$\text{MH-CHISQ} = \frac{(|\sum_j A_j - \sum_j E(A_j)| - \frac{1}{2})^2}{\sum_j \text{Var}(A_j)} \quad . \tag{11}$$

It may be shown, for example Birch (1964) or Cox (1970), that a test based on MH-CHISQ is the uniformly most powerful unbiased test of H_0 versus H_1. Hence no other test can have higher power somewhere in H_1 than the one based on MH-CHISQ unless the other test violates the size constraint on the null hypothesis or has lower power than the test's size somewhere else on H_1. Under H_0, MH-CHISQ has an approximate chi-square distribution with one degree of freedom. It corresponds to the single degree of freedom chi-square test given by Mellenberg (1982) for testing no "bias" against the hypothesis of "uniform bias." It is not identical to the test proposed by Mellenberg but in many practical situations they give virtually identical results even though Mellenberg's proposal involves an iterative log-linear model fitting process. The MH procedure is not iterative.

Mantel and Haenszel also provide an estimate of α, the common odds-ratio across the 2×2 tables. Their estimator is given by

$$\hat{\alpha}_{MH} = \frac{\sum A_j D_j/T_j}{\sum B_j C_j/T_j} \quad . \tag{12}$$

The odds-ratio is on the scale of 0 to ∞ with α=1 playing the role of a null value of no dif. It is convenient to take logs of $\hat{\alpha}_{MH}$ to put it into a symmetric scale in which 0 is the null value. Thus we have proposed that

$$\hat{\Delta}_{MH} = -\frac{4}{1.7} \ln(\hat{\alpha}_{MH}) = -2.35 \ln(\hat{\alpha}_{MH}) \tag{13}$$

be used as a measure of the amount of dif. $\hat{\Delta}_{MH}$ has the interpretation of being a measure of dif in the scale of differences in item difficulty as measured in the ETS "delta scale," (Holland and Thayer, 1985).

When using $\hat{\alpha}_{MH}$ or $\hat{\Delta}_{MH}$ it is useful to have a simple interpretation of these values. The value of $\hat{\alpha}_{MH}$ is the average factor by which the odds that a member of R is correct on the studied item exceeds the corresponding odds for a comparable member of F. Values of $\hat{\alpha}_{MH}$ that exceed 1 correspond to items on which the reference group performed better on average than did comparable members of the focal group. The value of $\hat{\Delta}_{MH}$ is the average amount more difficult that a member of R found the studied item than did comparable members of F. Values of $\hat{\Delta}_{MH}$ that are negative correspond to items that the reference group found easier on average than did comparable focal group members. The parameters, α and $\ln(\alpha)$, are also called "partial association" parameters because they are analogous to the partial correlations used with continuous data. The matching variable is "partialled out" of the association between group membership and performance on the studied item, (Birch, 1964).

Mantel and Haenszel proposed both the test statistic MH-CHISQ and the parameter estimate $\hat{\alpha}_{MH}$. Since that initial work many authors have contributed to the study of these procedures; the main results are as follows.

(a) The effect of the continuity correction is to improve the calculation of the observed significance levels using the chi-square table rather than to make the size of the test equal to the nominal value. Hence simulation studies routinely find that the actual size of a test based on MH-CHISQ is smaller than the nominal value. However, the observed significance level of a large value of MH-CHISQ is better approximated by referring MH-CHISQ to the chi-square tables than by referring the expression in (8) to these

tables. The continuity correction is simply to improve the approximation of a discrete distribution (i.e., MH-CHISQ) by a continuous distribution (i.e., one degree-of-freedom chi-square).

(b) $\hat{\alpha}_{MH}$ is a consistent estimator of the α in (8) and the variability of $\hat{\alpha}_{MH}$ is nearly optimal over the range $\frac{1}{3} < \alpha < 3$ which translates into $-2.6 < \Delta < 2.6$ under the log transformation in (13). Outside this range $\hat{\alpha}_{MH}$ or $\hat{\Delta}_{MH}$ are still reasonably efficient, but very large (or small) values of α are not as accurately estimated by $\hat{\alpha}_{MH}$ as they are by maximum likelihood. Since larger values of α are easy to detect using MH-CHISQ, this is not an important limitation.

(c) Standard error formulas for $\hat{\alpha}_{MH}$ and $\hat{\Delta}_{MH}$ that work in a variety of circumstances have taken a long time to develop. Important contributions have been Hauck (1979), Breslow and Laing (1982), and Flanders (1985). Recent joint work with A. Phillips suggests that the following approximate variance formula for $\ln(\hat{\alpha}_{MH})$ is valid whenever the numerator and denominator of $\hat{\alpha}_{MH}$ are both large:

$$\mathrm{Var}(\ln(\hat{\alpha}_{MH})) = \frac{1}{2U^2} \sum_j \left[T_j^{-2}(A_j D_j + \hat{\alpha}_{MH} B_j C_j)(A_j + D_j + \hat{\alpha}_{MH}(B_j + C_j))\right], \quad (14)$$

where

$$U = \sum_j A_j D_j / T_j.$$

This approximate variance formula agrees with well-known variance estimates for $\ln(\hat{\alpha}_{MH})$ in the few cases in which these are available. It is discussed more extensively in Phillips and Holland (1986).

It is sometimes helpful to show how $\hat{\alpha}_{MH}$ is expressed as a weighted average of the sample cross-product ratios in each of the K 2×2 tables. These are the values

$$\hat{\alpha}_j = \frac{A_j D_j}{B_j C_j} . \tag{15}$$

Hence

$$\hat{\alpha}_{MH} = \frac{\Sigma w_j \hat{\alpha}_j}{\Sigma w_j} \tag{16}$$

where

$$w_j = B_j C_j / T_j .$$

In their discussion of chi-square techniques, Marascuilo and Slaughter consider Cochran's (1954) test. In this test, instead of using the odds-ratio in each table as a measure of dif in the j^{th} matched group, the difference in proportion is used, i.e.

$$\frac{A_j}{n_{Rj}} - \frac{C_j}{n_{Fj}} . \tag{17}$$

These are averaged together with the weights,

$$n_{Rj} n_{Fj} / T_j ,$$

to get an overall average difference across all matched groups. More recently, Dorans and Kulick (1986) have suggested applying the weights n_{Fj} to the difference in (17) to get an overall standardized measure of dif for the item. Dorans and Kulick do not develop a test based on their measures, but it is evident that such a test, similar to Cochran's test, could be developed. Since Dorans and Kulick are primarily interested in a good descriptive measure of dif their choice of weights does not correspond to a statistically optimal test of H_0.

In summary, the Mantel-Haenszel procedure is a natural extension of the ideas behind the chi-square procedures of Scheuneman and others. It provides a single degree-of-freedom chi-square test that is powerful against realistic

alternatives to H_0, it allows detailed and careful matching on relevant criteria, and it provides a single summary measure of the magnitude of the departure from H_0 exhibited by the studied item.

4. THE MH PROCEDURE AND IRT MODELS

It is generally believed that there is, at best, only a rough correspondence between the "chi-square" types of procedures for studying dif and the more "theoretically preferred" methods based on item response theory (IRT). For examples of this view see Scheuneman (1979), Marascuilo and Slaughter (1981) and Shepard, Camilli, and Williams (1984). In this section we show that the MH procedure highlights a close connection between these two important classes of procedures. Our observations on this point are stongly influenced by the work of our colleague, Paul Rosenbaum -- see Rosenbaum (1985, 1986).

We adopt the notation and terminology for discussing IRT models given in Holland (1981), Cressie and Holland (1983), Rosenbaum (1984), and Holland and Rosenbaum (1986). Thus x_k is the 0/1 indicator of a correct response in item k, k=1,..., J, and $x = (x_1, \ldots, x_J)$ denotes a generic response vector -- there are 2^K possible values of x. In any population of examinees we let p(x) denote the proportion of them who would produce the response vector x if tested. Then

$$p(x) \geq 0, \quad \sum_x p(x) = 1.$$

An IRT model assumes that the value of p(x) is specified by an equation of the form

$$p(x) = \int \left[\prod_{k=1}^{J} P_k(\theta)^{x_k} Q_k(\theta)^{1-x_k} \right] dG(\theta). \qquad (18)$$

In (18), $P_k(\theta) = 1 - Q_k(\theta)$ is the item characteristic curve (ICC) for item k and $G(\theta)$ is the distribution function of the latent ability, θ, across the population of examinees.

It is customary to restrict the ICCs and θ in various ways. For example, θ is usually a scaler (not a vector) and the P_k are assumed to be monotone increasing functions of θ. Holland and Rosenbaum (1986) point out that without some restriction to this type IRT models are vacuous. Parametric assumptions such as the 1-, 2-, or 3-parameter logistic form for $P_k(\theta)$ may also be imposed.

If there are two subpopulations of examinees, R and F, then there are corresponding values $p_R(\mathbf{x})$ and $p_F(\mathbf{x})$. In general, each subpopulation will have its own ICCs, i.e.

$$P_{kR}(\theta) \text{ and } P_{kF}(\theta) \quad k=1,\ldots,J$$

as well as its own ability distribution,

$$G_R(\theta) \text{ and } G_F(\theta).$$

Lord (1980) states the hypothesis of no <u>dif</u> in terms of an IRT model. For item k it is

$$H_{0(IRT)} : P_{kR}(\theta) = P_{kF}(\theta) = P_k(\theta) \quad \text{for all } \theta.$$

Thus, if $H_{0(IRT)}$ holds for all k then $p_R(\mathbf{x})$ and $p_F(\mathbf{x})$ have the representations:

$$p_R(\mathbf{x}) = \int \left[\prod_{k=1}^{J} P_k(\theta)^{x_k} Q_k(\theta)^{1-x_k} \right] dG_R(\theta)$$

$$p_F(\mathbf{x}) = \int \left[\prod_{k=1}^{J} P_k(\theta)^{x_k} Q_k(\theta)^{1-x_k} \right] dG_F(\theta). \qquad (19)$$

Rosenbaum (1985) considers tests of the hypothesis that a representation like (19) exists for $p_R(\mathbf{x})$ and $p_F(\mathbf{x})$ in which R has a "higher" distribution of θ than does F.

The integrals in (18) and (19) are not easy to work with except in one special case, i.e., the Rasch model. For this model $P_k(\theta)$ has the logistic form

$$P_k(\theta) = e^{\theta - b_k}/(1 + e^{\theta - b_k}). \tag{20}$$

If (20) is inserted into (18) then Holland and Cressie (1983) show that p(x) may be expressed as

$$p(x) = p(0)\left[\prod_{k=1}^{J} f_k^{x_k}\right]\mu(x_+) \tag{21}$$

where

$$f_k = e^{-b_k}, \tag{22}$$

$$x_+ = \sum_k x_k$$

and

$$\mu(t) = E(U^t) \quad t=0,1,\ldots,J. \tag{23}$$

In (23), U is a positive random variable whose distribution depends on the ICCs and on the ability distribution $G(\theta)$. Hence if we apply (21) to $p_R(x)$ and $p_F(x)$ without assuming $H_{0(IRT)}$ we get

$$p_R(x) = p_R(0)\left[\prod_{k=1}^{J} f_{kR}^{x_k}\right]\mu_R(x_+) \tag{24}$$

and

$$p_F(x) = p_F(0)\left[\prod_{k=1}^{J} f_{kF}^{x_k}\right]\mu_F(x_+). \tag{25}$$

Now suppose that we wish to apply the MH procedure in this situation and that we take as the matching variable the total score on the test X_+. If item 1 is the studied item then the relevant population probabilities for R are of the form

$$p_{Rj} = P(X_1 = 1 | X_+ = j, R).$$

Using (24) this can be expressed as

$$p_{Rj} = \frac{P(X_1 = 1,\ X_+ = j)}{P(X_+ = i)} = \frac{f_{1R}\ S_{J-1,j-1}(f_R^*)}{S_{J,j}\ (f_R)} \tag{26}$$

where

$$f_R = (f_{1R},\ldots,f_{JR}) = (f_{1R},\ f_R^*)$$

and

$$S_{J,j}(\mathbf{f}) = \sum_{\mathbf{x}\ :\ x_+=j} \left[\prod_{k=1}^{J} f_k^{x_k}\right]$$

(i.e., the symmetric function of J-variables of degree j).

Similarly,

$$q_{Rj} = \frac{S_{J-1,j}(f_R^*)}{S_{J,j}(f_R)} \quad . \tag{27}$$

Hence the odds for success on item 1 in R in the $j^{\underline{th}}$ matched set are:

$$\frac{p_{Rj}}{q_{Rj}} = f_{1R}\,\frac{S_{J-1,j-1}(f_R^*)}{S_{J-1,j}(f_R^*)} \tag{28}$$

Similar equations hold for p_{Fj} and q_{Fj} and the corresponding odds are

$$\frac{p_{Fj}}{q_{Fj}} = f_{1F}\,\frac{S_{J-1,j-1}(f_F^*)}{S_{J-1,j}(f_F^*)} \quad . \tag{29}$$

Now suppose that for items 2 through J there is no <u>dif</u>, i.e.,

$$f_{kF} = f_{kR} \qquad k=2,\ldots,J,$$

so that $$f_F^* = f_R^*.$$

Then the population odds-ratio in each 2×2 table is

$$\frac{p_{Rj}}{q_{Rj}} \Big/ \frac{p_{Fj}}{q_{Fj}} = \frac{f_{1R}}{f_{1F}} = e^{b_{1F} - b_{1R}} \tag{30}$$

Equation (30) is a statement of H_1 in (6) with $\alpha = e^{b_{1F} - b_{1R}}$, so that for the Rasch model the hypothesis for which the MH procedure was developed holds exactly in the population under the following conditions.

(a) The items 2, 3, ..., J exhibit no dif, but the studied item may exhibit dif,

(b) the criterion for matching, X_+, includes the studied item,

(c) the data are random samples from R and F.

This result is a little surprising since the inclusion of the studied item in the criterion seems to go against the traditional uses of the MH procedures in medical applications. However, it can be shown that if the studied item is excluded from the criterion then the null hypothesis H_0 is not satisfied even though $H_{0(IRT)}$ is satisfied for every item.

For example, when the criterion for matching is $X_* = \sum_{k\geq 2} X_k$ and item 1 is the studied item, the relevant population probabilities are

$$p_{Rj} = P(X_1 = 1 \mid X_* = j, R)$$

and

$$p_{Fj} = P(X_1 = 1 \mid X_* = j, F).$$

It is easy to show that the equations that corresponds to (28) and (29) are

$$\frac{p_{Rj}}{q_{Rj}} = f_{1R} \frac{\mu_R(j+1)}{\mu_R(j)} \tag{31}$$

and

$$\frac{p_{Fj}}{q_{Fj}} = f_{1F} \frac{\mu_F(j+1)}{\mu_F(j)} . \tag{32}$$

Hence the odds-ratio in the $j^{\underline{th}}$ matched set is

$$\alpha_j = \frac{f_{1R}}{f_{1F}} \left[\frac{\mu_R(j+1)}{\mu_R(j)} \Big/ \frac{\mu_F(j+1)}{\mu_F(j)} \right]. \tag{33}$$

Thus the α_j are not constant across the 2×2 tables. The ratio of moments in (33) is related to order relationships between the distributions of the random

variable U, from (23), in R and F. For example, if the distribution of U for F is "lower" than that for R then we will have

$$\frac{\mu_R(j+1)}{\mu_R(j)} \Big/ \frac{\mu_F(j+1)}{\mu_F(j)} \geq 1, \text{ for all } j = 1, 2, \ldots .$$

This analysis raises the issue of whether the studied item should be included or not in the matching criterion. If it is not included, then the MH procedure will not behave correctly when there is no dif according to an IRT model. However the Rasch model analysis suggests that the inclusion of the studied item in the matching criterion does not mask the existence of dif, rather it is the inclusion of other items exhibiting dif in the criterion that could lead to the finding that no dif exists for the studied item when in fact it does. This idea leads to two steps.

Step 1: Purify the matching criterion by eliminating items based on a preliminary dif or impact analysis (Kok et al, 1985, make a similar suggestion).

Step 2: Use as the matching criterion the total score on all items left in the purified criterion plus the studied item -- even if it is then omitted from the criterion of all other items when they are studied in turned.

It is possible that we have drawn too heavily on the analysis of the Rasch model and a good deal of simulation work may be necessary before we know for sure if our suggestions hold in greater generality. We have begun some of that work and will report on it later. However, to date the results of the simulation study corraborates our proposal regarding the inclusion or exclusion of the studied item in the criterion.

The issue of including and excluding an item from the criterion shows the need for making these adjustments in the computational formulas for $\hat{\alpha}_{MH}$ and MH-CHISQ. These are as follows.

If the K 2×2 tables have been assembled for a number right score S as the matching criterion that does not include the studied item and we wish to include it in the score, then the 2×2 tables need to be altered to these.

Score on Studied Item

	1	0	Total
R	A_{j-1}	B_j	n'_{Rj}
F	C_{j-1}	D_j	n'_{Fj}
	m_{1j-1}	m_{0j}	T'_j

The values of MH-CHISQ and $\hat{\alpha}_{MH}$ are then computed from these tables.

Similarly, if S contains the score of the studied item and we wish to eliminate it this is done by using these 2×2 tables.

Score on Studied Item

	1	0	Total
R	A_{j+1}	B_j	n''_{Rj}
F	C_{j+1}	D_j	n''_{Fj}
	m_{1j+1}	m_{0j}	T''_j

Thus it is a simple matter to compute either $\hat{\alpha}_{MH}$ or MH-CHISQ including or excluding the studied item from a number-right-score matching criterion. If the matching criterion is a formula-score or a grouped, number-right-score then it is not easy to adjust for the inclusion of the studied item into the criterion,

without recalculating the entire set of 2×2 tables.

5. DISCUSSION

There are many procedures that have been proposed for the study of dif over the last twenty years, and the introduction of a new one, associated with names that are unfamiliar to psychometricians, is likely to be regarded skeptically. However, we have tried to show that the MH procedure, drawn from the field of biostatistics, fits squarely into the network of ideas developed by previous workers in the field of "item bias". In addition, standard statistical concepts, such as tests, hypotheses, error of type I and II, estimates, and standard errors all fit neatly into the package. Connections between chi-square methods and IRT based methods are made evident by studying the Mantel-Haenszel procedure.

We believe that the view that IRT based approaches to dif are "theoretically preferred" over chi-square based procedures is not a very precise way of describing the situation. It is certainly true that likelihood ratio tests of $H_{0(IRT)}$ in the context of specific parametric IRT models (i.e., 3PL ICCs and Normal θ-distributions) are statistically optimal (or very nearly so) in the sense of power and efficiency when these models actually hold. If the data really are generated by such models, as they would be in a simulation, then no other test of the equality of two ICCs for the same item, at the given significance level can have larger power than these likelihood ratio tests. However, it is only the procedures based on marginal maximum likelihood, as advocated by Bock and Aitkin (1981), that can yield true likelihood ratio tests (e.g., see Thissen, Wainer, and Steinberg (1985).) IRT-based procedures that depend on multiple LOGIST calibrations do not automatically result in tests of $H_{0(IRT)}$ and estimates of ICC differences that are optimal. Furthermore, even the marginal

maximum likelihood procedures are not optimal when the assumed model is wrong.

In our view, parametric IRT models provide an important testing ground for evaluating dif procedures. Under $H_{0(IRT)}$, test statistics ought to achieve significance levels that are close to the nominal values regardless of the choices of G_R, G_F, and the ICCs, $P_k(\theta)$. Against alternatives to $H_{0(IRT)}$, the likelihood ratio procedure will set the upper bounds on the power and efficieny of any test procedure, including the LOGIST-based procedures or chi-square procedures like the Mantel-Haenszel. Our use of a specific IRT model (the Rasch model) to evaluate the Mantel-Haenszel procedure resulted in a new conception of the importance of including or excluding the studied item in the criterion. This shows the advantage of a theoretical analysis. We were led to that analysis by the empirical finding that including the studied item in the test score used as a matching criterion had a measurable and consistent effect on the values of $\hat{\alpha}_{MH}$ and $\hat{\Delta}_{MH}$ computed in real data. The $\hat{\Delta}_{MH}$ values will shift by an amount that is nearly independent of the studied item but which did depend on the overall differences in performance on the criterion test between R and F. The bigger the difference the bigger the shift. This is exactly what is predicted by equation (33) when there are large differences between the θ-distributions, G_F and G_R.

Our conjecture is that it is correct to include the studied item in the matching criterion when it is being analysed for dif, but if it has substantial dif then that item should be excluded from the criterion used to match examinees for any other studied item. The first "inclusion" is to control the size of the test given by MH-CHISQ while the second "exclusion" is to prevent large dif items from degrading the power of this test. Such an approach is independent of the MH procedure and can be incorporated into other chi-square techniques, or

into the iterative logit technique discussed by Kok et al. (1985).

A final note on costs. The MH procedure is very inexpensive to use compared to IRT analyses. For example, runs that involve 50 items and 2500 examinees cost about $10 on a typical mainframe computer. Our main reason for pursuing this approach has been to provide ETS with a practical, and yet powerful tool for the study of dif that incorporates all of the advances in methodology that have occurred since the late 1970s.

REFERENCES

Angoff, W. H. and Ford, S. F. (1973) Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-106.

Baker, F. B. (1981) A criticism of Scheuneman's item bias technique. Journal of Educational Measurement, 18, 59-62.

Berk, R. A. (ed.) (1982) Handbook of Methods for Detecting Test Bias. Baltimore, MD: Johns Hopkins University Press.

Birch, M. W. (1964) The detection of partial association, I: The 2×2 case. Journal of the Royal Statistical Society, Series B, 26, 313-324.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975) Discrete Multivariate Analysis: Theory and Practice. Cambridge, MA: MIT Press.

Bock, R. D. and Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: Application of the EM algorithm. Psychometrika, 46, 443-460.

Breslow, N. E. and Liang, K. Y. (1982) The variance of the Mantel-Haenszel Estimator. Biometrics, 38, 943-952.

Cardall, C. and Coffman, W. E. (1964) A method for comparing the performance of different groups on the items in a test. Princeton, NJ: Educational Testing Service, Research Bulletin RB-64-61.

Cochran, W. G. (1954) Some methods for strengthening the common χ^2 test. Biometrics, 10, 417-451.

Cox, D. R. (1970) Analysis of Binary Data. London: Methuen and Co., Ltd.

Cressie, N. and Holland, P. W. (1983) Characterizing the manifest probabilities of latent trait models. Psychometrika, 48, 129-141.

Dorans, N. J. and Kulick, E. M. (1986) Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, in press.

Flanders, W. D. (1985) A new variance estimator for the Mantel-Haenszel odds-ratio. Biometrics, 41, 637-642.

Hauck, W. W. (1979) The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. Biometrics, 35, 817-819.

Holland, P. W. (1981) When are item response models consistent with observed data? Psychometrika, 46, 79-92.

Holland, P. W. (1985) On the study of Differential Item Performance without IRT. Proceedings of the Military Testing Association, October 1985, in press.

Holland, P. W. and Phillips, A. (1986) A new estimator of the variance of the Mantel-Haenszel log-odds-ratio estimator. submittted to Biometrics.

Holland, P. W. and Rosenbaum, P. R. (1986) Conditional association and unidimensionality in monotone latent variable models. Annals of Statistics, in press.

Holland, P. W. and Thayer, D. T. (1985) An alternative definition of the ETS delta scale of item difficulty. Princeton, NJ: Educational Testing Service, Research Report RR-85-43.

Kok, F. G., Mellenbergh, G. J., and van der Flier, H. (1985) Detecting experimentally induced item bias using the iterative logit method. Journal of Educational Measurement, 22, 295-303.

Lord, F. M. (1980) Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum.

Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Marascuilo, L. A. and Slaughter, R. E. (1981) Statistical procedures for identifying possible sources of item bias based on χ^2 statistics. Journal of Educational Measurement, 18, 229-248.

Mellenberg, G. J. (1982) Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118.

Rosenbaum, P. R. (1984) Testing the conditional independence and monotonicity assumptions of item response theory. Psychometrika, 49, 425-435.

Rosenbaum, P. R. (1985) Comparing distributions of item responses for two groups. British Journal of Mathematical and Statistical Psychology, 38, 206-215.

Rosenbaum, P. R. (1986) Comparing item characteristic curves. (manuscript).

Scheuneman, J. D. (1979) A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Scheuneman, J. D. (1981) A response to Baker's criticism. Journal of Educational Measurement, 18, 63-66.

Shepard, L. A., Camilli, G., and Williams, D. M. (1984) Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Thissen, D., Wainer, H., and Steinberg, L. (1985) Studying differential item performance with item response theory. Paper presented at the Military Testing Association meeting, San Diego, California, October 21, 1985.